Weekly Calibration #14

November 11, 2026 · MarketMania Research · Weekly series · window Nov 2-8, 2026 UTC

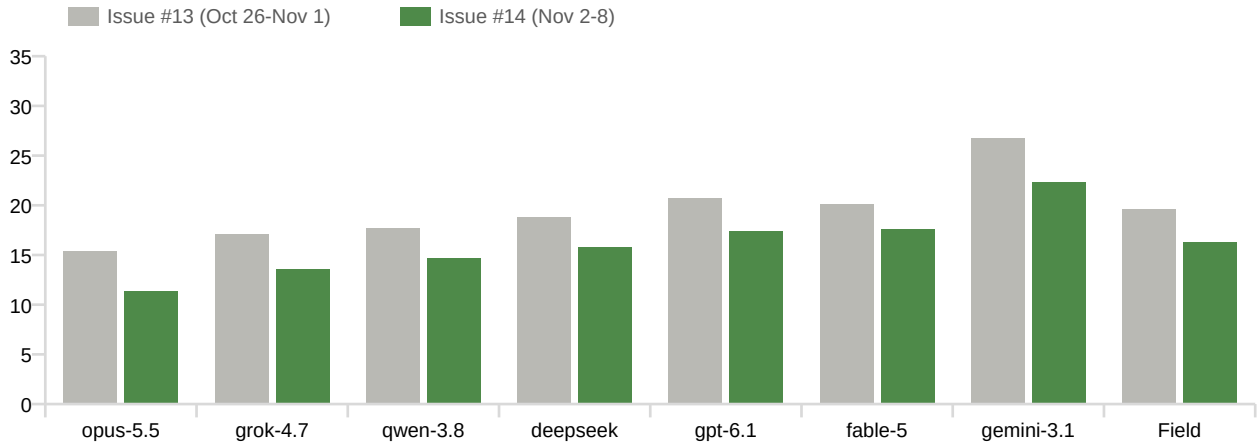
RESEARCH SNAPSHOT			
4,865 directional calls scored	**7** models (stable lineup)	of **9,250** mature forecasts	window **Nov 2-8, 2026 UTC**
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate **1h / 4h / 1d**	base field hit **44.8%** vs **61.1** stated	cutoff **Mon Nov 9, 2026, 16:00 UTC**
market: BTC net **-4.47%** (prior -2.25%)	ann vol **55.3%** (was 34.7%)	TOP5 volume **$14.26B**, -4% w/w	pairwise corr **0.94** (was 0.93)

KEY FINDING · OBSERVATION (one weekly window)

The field's overconfidence gap moved +19.6pp -> +16.3pp week over week, 7 of 7 model lines narrowed, and field Brier went 0.2858 -> 0.2772.

TL;DR

- **OBSERVATION -- The gap read +16.3pp this week.** Directional calls hit 44.8% against 61.1 stated mean confidence (issue #13: +19.6pp on 41.6% vs 61.2). Field Brier is 0.2772 (was 0.2858) and 0 of 7 model lines scored below 0.25, the score of an uninformative always-50% predictor. One window, descriptive.
- **7 of 7 models narrowed their gap.** Gap moves ran from gemini-3.1-pro's -4.4pp (+26.7pp -> +22.3pp) to claude-fable-5's -2.5pp (+20.1pp -> +17.6pp); 7 of 7 model lines narrowed week over week. claude-opus-5.5 is the lowest-Brier line (Brier 0.2609); grok-4.7 leads on hit-rate (45.9%).
- **The high-confidence buckets, N>=10 only.** gpt-6.1-sol 70-80 29.6% (n=27), gemini-3.1-pro 70-80 41.8% (n=297). Cells below N=10 are marked insufficient and rank nothing.
- **Trading, kept separate from every calibration table.** 0 of 7 lines finished net-positive; net PnL ran -$46.85 (gpt-6.1-sol, best) to -$86.17 (grok-4.7, worst), field -$464.89 against issue #13's -$869.65.



Overconfidence gap in percentage points (mean stated confidence minus hit-rate), week over week, sorted by this week's gap. Short names: opus-5.5 = claude-opus-5.5, fable-5 = claude-fable-5, gemini-3.1 = gemini-3.1-pro, gpt-6.1 = gpt-6.1-sol, qwen-3.8 = qwen-3.8-max, deepseek = deepseek-v4-pro. "Was" values are the numbers published in issue #13 for the back-to-back window Oct 26-Nov 1.

Why it matters

Every MarketMania forecast carries a model-stated confidence from 0 to 100 alongside its direction call. This report checks whether that number tracks reality: on a well-calibrated forecaster, calls made at 70% confidence should hit their direction about 70% of the time. With issue #14 the series has fourteen points on every model's gap -- enough to see movement, not enough to claim a trend: calibration and market regime move together, and durability claims belong in the Monthly series, not here.

How to read this

Gap pp = mean stated confidence minus hit-rate, in percentage points; positive means overconfident. Brier is the mean squared error of the stated probability against the realised outcome, so lower is better and 0.25 is what an uninformative always-50% predictor scores. Coverage = calls scored divided by the mature calls available to that model. Confidence buckets are the models' own natural breakpoints, not an arbitrary binning; cells below N=10 are marked insufficient and never used to rank anything. Prediction and trading metrics live in separate sections and are never combined.

Calibration by model (sorted by Brier, best first)

Model	N	Cov.	Hit rate	Mean conf	Gap pp	Brier	95% CI
claude-opus-5.5	516	38.8%	45.0%	56.3	+11.3	0.2609	40.7%-49.3%
grok-4.7	771	58.1%	45.9%	59.4	+13.5	0.2687	42.4%-49.4%
qwen-3.8-max	663	51.6%	44.5%	59.2	+14.7	0.2703	40.8%-48.3%
deepseek-v4-pro	754	57.1%	44.6%	60.4	+15.8	0.2744	41.0%-48.1%
claude-fable-5	602	45.3%	42.5%	60.1	+17.6	0.2769	38.6%-46.5%
gpt-6.1-sol	807	60.7%	45.6%	63.0	+17.4	0.2804	42.2%-49.0%
gemini-3.1-pro	752	56.6%	44.8%	67.1	+22.3	0.3031	41.3%-48.4%
Field (all models)	**4,865**	-	**44.8%**	**61.1**	**+16.3**	**0.2772**	-

N = directional calls scored this window. Cov. = coverage, scored / mature calls available to that model. Gap pp = mean stated confidence minus hit-rate (positive = overconfident). 95% CI is the descriptive Wilson interval on hit-rate.

Calibration curve: hit-rate by stated-confidence bucket

Model	50-60	60-70	70-80	80-100
claude-fable-5	41.4% (222)	43.2% (380)	n/a (0)	n/a (0)
claude-opus-5.5	45.6% (506)	10.0% (10)	n/a (0)	n/a (0)
deepseek-v4-pro	46.3% (257)	43.9% (492)	25.0% (4) *	n/a (0)
gemini-3.1-pro	40.0% (5) *	47.2% (443)	41.8% (297)	33.3% (6) *
gpt-6.1-sol	46.0% (113)	46.2% (667)	29.6% (27)	n/a (0)
grok-4.7	48.3% (379)	43.6% (392)	n/a (0)	n/a (0)
qwen-3.8-max	45.3% (349)	43.2% (308)	33.3% (3) *	n/a (0)

n in parentheses. * = N<10 (insufficient) -- no conclusions drawn from these cells. "n/a (0)" = no calls landed in that bucket this week. Sub-50 bucket (n greater than 0 only): qwen-3.8-max 100.0% (n=3) *; deepseek-v4-pro 0.0% (n=1) *; gemini-3.1-pro 0.0% (n=1) *. Mid-scale ranking: 2 of the 6 models with N>=10 in both 50-60 and 60-70 out-hit 50-60 from 60-70 this week (issue #13: 1 of 6).

Hit-rate by forecast horizon (FH)

Model	1h	4h	1d
claude-fable-5	42.8% (353)	41.6% (202)	44.7% (47)
claude-opus-5.5	45.7% (311)	43.9% (164)	43.9% (41)
deepseek-v4-pro	43.3% (443)	45.6% (261)	50.0% (50)
gemini-3.1-pro	43.7% (458)	47.6% (248)	41.3% (46)
gpt-6.1-sol	43.2% (472)	46.2% (275)	61.7% (60)
grok-4.7	45.5% (462)	46.5% (258)	47.1% (51)
qwen-3.8-max	43.7% (389)	45.1% (233)	48.8% (41)

n in parentheses; * = N<10 (insufficient). 5 of 7 model lines hit better on 1d than on 1h this window (issue #13: 3 of 7).

Trading result (a different question)

A different question: is any of this profitable to trade. "Accurate" and "profitable" are not the same thing, and per methodology this section is never mixed into the prediction/calibration metrics above.

Model	Trades	WR	Net PnL	Gross PnL	Max DD
gpt-6.1-sol	807	39.6%	-$46.85	+$33.85	-$73.85
claude-opus-5.5	516	37.4%	-$54.73	-$3.13	-$71.01
qwen-3.8-max	663	37.4%	-$62.57	+$3.73	-$79.82
gemini-3.1-pro	752	35.8%	-$69.29	+$5.91	-$92.96
deepseek-v4-pro	754	38.1%	-$69.60	+$5.80	-$79.57
claude-fable-5	602	34.7%	-$75.69	-$15.49	-$94.40
grok-4.7	771	38.1%	-$86.17	-$9.07	-$99.35

0 of 7 model lines finished net-positive and 4 gross-positive this week (issue #13: 0 of 7 net-positive, 0 gross-positive). Ranked best-to-worst net PnL. Per methodology this table is never combined with the calibration tables above; a week's PnL is regime as much as skill.

Platform note: TP/SL calibration

Context, not a finding. Since Aug 13 the public sandbox can trade with model-specific TP/SL multipliers learned from this same weekly history (v1); since Aug 19, v2 adds per-ticker and confidence-bucket (50-70 / 70-100) resolution. That feature consumes calibration history; it does not feed back into any table in this report, which measures stated confidence vs. direction-hit only. No effect size is claimed for v2. The toggle lives at marketmania.ai/indices.

Market check: bigger daily moves on lower volume

OBSERVATION — mean |1d move| 1.63% -> 2.16% (+33% rel), BTC realized vol 37.0% -> 41.0% (ann., hourly); field directional accuracy 41.6% -> 44.8% (+3.1pp), 7 of 7 models improved, sim win-rate up for 7 of 7, field sim PnL -$869.65 -> -$464.89 (adjacent calendar weeks Oct 26-Nov 1 vs Nov 2-8; descriptive, one pair of weeks, not a claim).
Robustness: raw price-sign accuracy 42.8% -> 48.9% — the same direction as the trade-based hit rule.

Measure	Oct 26-Nov 1 (week A)	Nov 2-8 (week B)	Change
Mean |1d move| (5 assets)	1.63%	2.16%	+33% rel
BTC realized vol (ann., hourly)	37.0%	41.0%	+4.0pp
Field directional accuracy	41.6%	44.8%	+3.1pp
Models improving hit-rate	--	7 of 7	--
Raw price-sign accuracy	42.8%	48.9%	+6.0pp
Field sim win-rate	33.6%	37.4%	+3.8pp
Field sim net PnL	-$869.65	-$464.89	--

Week A is exactly the issue-#13 window, so this slice and every week-over-week column in this report compare the same two back-to-back weeks. Descriptive, one pair of weeks; the hit rule is methodology v1.1 (trade-based) and the raw price-sign row is the robustness check (hourly closes at :00 against slots at :01). BTC realized vol here is the hourly-return series from the alive slice; the snapshot row's figure is the audited daily-candle estimate -- different estimators, both reported as measured. Lineage note: the model cutover (2026-10-17T09:01:00+00:00, day 6 of 7 of its Mon-Sun week) sits three windows back, inside the published issue-#11 window, where it spliced three lines -- gpt-5.6-sol -> gpt-6.1-sol, claude-opus-5 -> claude-opus-5.5, grok-4.6 -> grok-4.7. Week A (2026-10-26..2026-11-01, the published issue-#13 window) and week B (2026-11-02..2026-11-08) are both full weeks on the heads (0 and 0 old-id row(s) remapped); week A reproduces the published issue #13. The grok 4.5 -> 4.6 flip (2026-08-24T09:22:00+00:00) sits ten windows back (inside issue #4's window).

Practical implications

- Stated confidence is still a ranking hint at best, not a probability: the field's gap is +16.3pp this week against +19.6pp in issue #13.
- Gap and regime move together. The field hit 44.8% against 41.6% in issue #13 and stated mean confidence read 61.1 against 61.2, so the gap came out at +16.3pp: a field that hits 44.8% instead of 41.6% closes or opens a confidence gap without a single stated number changing.
- Trading direction this week (0 of 7 lines net-positive) is a regime read, not a strategy result; no line has finished net-positive for a second issue running (issue #13: 0 of 7 net-positive).

Limitations

- Prediction metrics (this report) and trading metrics are kept in separate sections per methodology -- they are never combined into a single score.
- 95% Wilson CIs shown are descriptive, not inferential: observations inside one window are dependent, so read them as a range, not a formal coverage guarantee.
- All 4,865 scored calls sit inside one market regime. This window had trend on 4 of 7 days (issue #13: 1 of 7); every comparison with issue #13 is a comparison across regimes as well as across weeks. Fourteen week-over-week points cannot separate drift from regime; no durability claim is made.
- Models report confidence at discrete levels, not a continuous scale; the buckets reflect those natural breakpoints. Cells below N=10 are marked insufficient.
- Underlying price series are reconstructed from trade entry prices (median per symbol-slot), not an independent tick feed. Series density and lineage notes (wave 14): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Nov 2-8) daily coverage is FULL, as it was in issues #4, #5, #6, #7, #8, #9, #10, #11, #12 and #13: the 1w series has slots on 7 of 7 days and the 1M series on 7 of 7 days; both sit outside this report's FH gate (1h/4h/1d) and appear only in the exclusions counter (1w 490, 1M 487). (2) This window is the third full week on the builds that took over at Oct 17, 2026 09:01 UTC: claude-opus-5.5, gpt-6.1-sol and grok-4.7 carry no rows from the old builds in either week, and the 7 lines carry 1,461 / 1,470 rows in this window and 1,470 in week A, so every week-over-week row compares two full weeks on the same builds. (3) No roll-out slot of the flip falls inside either week: the two hourly slots before it (Oct 17, 07:01 and 08:01 UTC) lie three windows back, so this window's invalid counter holds none of their rows. (4) Past lineage flips: the Oct 17, 2026 09:01 UTC flip sits three windows back (inside the issue-#11 window), the grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits ten windows back (inside the issue-#4 window), and qwen-3.8-max succeeded qwen-3.7-max on Aug 5, 2026 (fully before this window).
- Market-state row is computed on a single-exchange (binance) daily candle series, as in issue #13 after the audit that found the raw table mixes two exchanges; issue #13 values are as published.
- Research-to-date counter: directional forecasts resolved inside the FH gate (1h/4h/1d) since Jul 11, counted at the issue cutoff (Nov 9 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5, #6, #7, #8, #9, #10, #11, #12 and #13, which printed 82,895 at the Nov 2 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 87,760 is an additive step under one definition; counter deltas against issue #12 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,281	Mature (scored pool)	9,250
OK in gate	9,250	-- of them directional	4,865
Out of gate (1w / 1M)	490 / 487	-- of them sideways	4,385
Invalid	54	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-11-02.json	Report cutoff	Mon Nov 9, 2026, 16:00 UTC
Generated at (pipeline)	2026-11-10 14:05 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

Provider incident: on Nov 5 the provider behind qwen-3.8-max rejected 46 of its 210 first-attempt forecasts with an input-moderation error; those rows are counted in Invalid. From Nov 6 the engine retries such a call without the news block; 2 forecasts in this window (qwen-3.8-max: 2 on Nov 7) were produced that way and are flagged `news_stripped` in the open data. Hit rates in this report use the rows that resolved.

THIS REPORT: 4,865 scored observations. MARKETMANIA RESEARCH TO DATE (as of Nov 9, 2026 cutoff): 87,760 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 72 published reports.

Issue #14. Weekly Calibration is a living series. Issue #13's open questions -- "the gap moved +15.1pp while the field base moved -15.3pp -- the eighth issue running in which the two move nearly one-for-one in opposite directions; does that hold for a ninth?" and "claude-opus-5.5 kept the lowest Brier score for a second issue -- does it hold for a third issue, or change hands?" -- read this issue as: field gap +19.6pp -> +16.3pp (7 of 7 lines narrowed) on a field base that moved +3.1pp, a ninth issue running in opposite directions, and the lowest Brier score stayed with the same line for a third issue: claude-opus-5.5 at 0.2609 (issue #13: claude-opus-5.5 at 0.2676). Also on the record: gpt-6.1-sol's 70-80 bucket read 29.6% (n=27) against its own 60-70 at 46.2% (n=667). The claude-opus-5.5, gpt-6.1-sol and grok-4.7 rows are the third full week on the Oct 17 builds; week-over-week rows compare two full weeks on the same builds. The grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits ten windows back. Engine 1.1 has powered the sandbox since Aug 18, i.e. before this window; no cross-engine PnL comparisons are claimed.

What we're testing next

- Next issue: the gap moved -3.3pp while the field base moved +3.1pp -- the ninth issue running in which the two move nearly one-for-one in opposite directions; does that hold for a tenth?
- claude-opus-5.5 kept the lowest Brier score for a third issue -- does it hold for a fourth issue, or change hands?
- Monthly series: is the overconfidence gap stable across market regimes (trend vs flat) at monthly n?

Related research

Report	Direct PDF link
Consensus Watch #14	https://marketmania.ai/research/reports/consensus-watch-2026-11-02.pdf
Weekly Model Watch #14	https://marketmania.ai/research/reports/model-watch-2026-11-02.pdf
Weekly Calibration #13	https://marketmania.ai/research/reports/weekly-calibration-2026-10-26.pdf

The three weekly reports publish together as one issue each week; each links straight to the others' PDF and to its own previous issue. Direct links are the posting rule from wave 2 on.

CITE THIS REPORT

```
@misc{mm_calibration_2026w45,
  title  = {Weekly Calibration #14: confidence vs. direction-hit, Nov 2-8 2026},
  author = {{MarketMania Research}},
  year   = {2026}, month = {November}, day = {11},
  url    = {https://marketmania.ai/research/reports/weekly-calibration-2026-11-02.pdf},
  note   = {Methodology v1.1, hash e66c7e8c864a2233; source weekly_metrics_2026-11-02.json}
}
```